hdidregress — Heterogeneous difference in differences

Description

`hdidregress` estimates average treatment effects on the treated (ATETs) that may vary over time and over treatment cohorts. Treatment cohorts are groups subject to treatment at different points in time. `hdidregress` provides four estimators: extended two-way fixed effects (TWFE), regression adjustment (RA), inverse-probability weighting (IPW), and augmented inverse-probability weighting (AIPW). See [CAUSAL] **teffects intro** for a discussion of RA, IPW, and AIPW estimators.

`hdidregress` is for repeated cross-sectional data. For panel data, see [CAUSAL] **xthdidregress**.

Quick start

Estimate ATETs of treatment `treat` on outcome `y` with group `grpvar` and time `tvar`; use the RA estimator and model `y` using covariate `x`

```
hdidregress ra (y x) (treat), group(grpvar) time(tvar)
```

Same as above, but use the TWFE estimator

```
hdidregress twfe (y x) (treat), group(grpvar) time(tvar)
```

Use the IPW estimator and model `treat` using covariate `z`

```
hdidregress ipw (y) (treat z), group(grpvar) time(tvar)
```

Use the AIPW estimator, model `y` using covariate `x`, and model `treat` using covariate `z`

```
hdidregress aipw (y x) (treat z), group(grpvar) time(tvar)
```

Same as above, but use the not-yet-treated group as the control group

```
hdidregress aipw (y x) (treat z), group(grpvar) time(tvar) ///
    controlgroup(notyet)
```

Same as above, but cluster at the `county` level

```
hdidregress aipw (y x) (treat z), group(grpvar) time(tvar) ///
    controlgroup(notyet) vce(cluster county)
```

Menu

Statistics > Causal inference/treatment effects > Continuous outcomes > Heterogeneous DID > Heterogeneous DID (TWFE)

Statistics > Causal inference/treatment effects > Continuous outcomes > Heterogeneous DID > Heterogeneous DID (RA)

Statistics > Causal inference/treatment effects > Continuous outcomes > Heterogeneous DID > Heterogeneous DID (IPW)

Statistics > Causal inference/treatment effects > Continuous outcomes > Heterogeneous DID > Heterogeneous DID (AIPW)

Syntax

Two-way fixed effects

```
hdidregress twfe (ovar [omvarlist]) (tvar) [if] [in] [weight],
    group(groupvar) time(timevar) [options]
```

Regression adjustment

```
hdidregress ra (ovar [omvarlist]) (tvar) [if] [in] [weight],
    group(groupvar) time(timevar) [options]
```

Inverse-probability weighting

```
hdidregress ipw (ovar) (tvar [tmvarlist]) [if] [in] [weight],
    group(groupvar) time(timevar) [options]
```

Augmented inverse-probability weighting

```
hdidregress aipw (ovar [omvarlist]) (tvar [tmvarlist]) [if] [in] [weight],
    group(groupvar) time(timevar) [options]
```

ovar is a continuous outcome of interest.

omvarlist specifies the covariates in the outcome model and may contain factor variables; see **[U] 11.4.3 Factor variables**.

tvar must be a binary variable indicating observations subject to treatment.

tmvarlist specifies the covariates in the treatment model and may contain factor variables; see **[U] 11.4.3 Factor variables**.

groupvar is a categorical variable that indicates the group level at which the treatment occurs.

timevar is a time variable.

options	Description
Model	
* `group(`*groupvar*`)`	specify group variable
* `time(`*timevar*`)`	specify time variable
`controlgroup(`*cgtype*`)`	specify the type of control group; default is `controlgroup(never)`
`cohortvar(`*cvar* [`, replace`]`)`	specify the variable name for the generated cohort
`usercohort(`*varname*`)`	specify name of cohort variable to be used for estimation
† `basetime(`*btspec*`)`	specify the type of base time for pretreatment periods; default is `basetime(adaptive)`
‡ `hettype(`*hetspec*`)`	specify the type of heterogeneity; default is `hettype(timecohort)`
SE/Robust	
`vce(`*vcetype*`)`	*vcetype* may be `cluster` *clustvar*, `robust`, `bootstrap`, or `jackknife`
Reporting	
`level(`#`)`	set confidence level; default is `level(95)`
[`no`]`log`	suppress iteration log
`nodots`	suppress replication dots
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
`coeflegend`	display legend instead of statistics

cgtype	Description
`never`	use the never-treated group as the control group; the default
`notyet`	use the not-yet-treated group as the control group

btspec	Description
`adaptive`	specify the adaptive base time for pretreatment ATETs; the default
`common`	specify a common base time for all pretreatment ATETs

hetspec	Description
`timecohort`	heterogeneous treatment effects over time and cohort; the default
`time`	heterogeneous treatment effects over time
`cohort`	heterogeneous treatment effects over cohort

*group() and time() are required.

†basetime() may be specified only when ra, ipw, or aipw is specified.

‡hettype() may be specified only when twfe is specified.

by, collect, and statsby are allowed; see [U] 11.1.10 Prefix commands.

fweights, aweights, and pweights are allowed; see [U] 11.1.6 weight.

coeflegend does not appear in the dialog box.

See [U] 20 Estimation and postestimation commands for more capabilities of estimation commands.

Options

Model

group(*groupvar*) specifies a group variable that indicates the group level at which the treatment occurs. *groupvar* may be, for example, states, counties, or hospitals. group() also defines the clusters for the default cluster–robust standard errors. group() is required. You may specify only one group variable.

time(*timevar*) specifies the time variable used to define treatment-time cohorts. time() is required.

controlgroup(*cgtype*) specifies the type of control group. A control group can be either a never-treated group or a not-yet-treated group. A never-treated group refers to the units that are untreated from the first to the last period. A not-yet-treated group refers to the units that are untreated up to a specific period. *cgtype* can be one of never, referring to the never-treated group, or notyet, referring to the not-yet-treated group. By default, *cgtype* is never.

cohortvar(*cvar* [, replace]) specifies the variable name *cvar* for the generated cohort variable. The cohort variable is a categorical variable indicating the period when the unit is first treated. By default, _did_cohort is used as the name of the cohort variable. If _did_cohort already exists in the dataset, it is replaced if option cohortvar() is not specified.

If suboption replace is specified, *cvar* is replaced.

usercohort(*varname*) specifies a variable to be used as a cohort indicator during estimation. By default, a cohort variable is generated using the information in the estimation sample to indicate the period when a unit is first treated. usercohort() overrides this default and allows you to provide a cohort indicator. This is useful, for instance, when there are gaps in the estimation sample, but you know a group was treated at the time when the gap is present in the data.

basetime(*btspec*) specifies how the base time is chosen when computing the pretreatment ATETs with the ra, ipw, or aipw estimator. *btspec* is one of adaptive (the default) or common.

adaptive specifies that the base time for pretreatment ATETs be chosen adaptively. The base time for each pretreatment period t for cohort g is the previous period, $t-1$.

common specifies that a common base time of $g-1$ be used for all pretreatment ATETs for cohort g. A long-run violation of the parallel trends assumption is easier to identify when using this common base time.

The base time for posttreatment periods is $g-1$, regardless of whether the adaptive or common base time is used for pretreatment periods.

hettype(*hetspec*) specifies time or cohort heterogeneity for the twfe estimator. By default, treatment is interacted with time and cohort. You may choose to keep one of time or cohort interactions using *hetspec*.

hetspec may be one of `timecohort` for heterogeneous treatment effects over both time and cohort, `time` for heterogeneous treatment effects over time only, or `cohort` for heterogeneous treatment effects over cohort only. By default, *hetspec* is `timecohort`.

SE/Robust

`vce(`*vcetype*`)` specifies the type of standard error reported, which includes types that allow for intragroup correlation (`cluster` *clustvar*), that are robust to intragroup correlation among group variable (`robust`), and that use bootstrap or jackknife sampling done at the individual level (`bootstrap`, `jackknife`); see [R] ***vce_option***.

`vce(cluster` *clustvar*`)`, the default, uses the variable specified in `group(`*groupvar*`)`.

Reporting

`level(#)`; see [R] **Estimation options**.

`log` and `nolog` specify whether to display the iteration log. The iteration log is displayed by default unless you used `set iterlog off` to suppress it; see `set iterlog` in [R] ***set iter***.

`nodots` suppresses display of the replication dots.

display_options: `noci`, `nopvalues`, `cformat(%`*fmt*`)`, `pformat(%`*fmt*`)`, `sformat(%`*fmt*`)`, and `nolstretch`; see [R] **Estimation options**.

The following option is available with `hdidregress` but is not shown in the dialog box:

`coeflegend`; see [R] **Estimation options**.

Remarks and examples

It is common to study the effects of a treatment, for example, a policy or intervention, on a group. `hdidregress` is for data where the treated groups are subject to the treatment at different points in time and they remain exposed to the treatment. For example, a health policy such as an increase in the age to purchase cigarettes is implemented in a given region, and over time, other regions decide to imitate the initiative. Another example is change in work policies across industries. Perhaps airlines implement a minimum number of hours between shifts for safety reasons. The policy is subsequently adopted by other similar industries. Some similar industries may never adopt the policy, remaining untreated, or it might be that all similar industries eventually adopt the policy.

`hdidregress` estimates ATET parameters that change over time and treatment cohorts (groups treated at different points in time). Each one of these ATETs has the same interpretation that the parameters of a two-time two-group difference-in-differences (DID) parameter would have. Because there are multiple DID parameters, we refer to them as heterogeneous treatment effects or as heterogeneous DID. This is in contrast to estimating only one ATET, which assumes there is no variation across time or cohort. If you assume no variation across time or cohort, you may use `didregress`; see [CAUSAL] **didregress**.

`hdidregress` provides four estimators: TWFE, outlined in Wooldridge (2021); RA, IPW, and AIPW, outlined in Callaway and Sant'Anna (2021). Each one of these estimators fits a model for the outcome of interest, a model for the treatment, or a model for both. For example, RA and TWFE model the outcome; IPW models the treatment; and AIPW models both. If the model for the outcome is correctly specified, RA and TWFE are best, with TWFE being more efficient. If the treatment model is correctly specified, IPW should be best. AIPW models both treatment and outcome. If at least one of the models is correctly

specified, it provides consistent estimates. Thus, it allows us to misspecify one of the models and still get consistent estimates, a property called double robustness. See [CAUSAL] **teffects intro** for a discussion of RA, IPW, and AIPW estimators.

`hdidregress` is for repeated cross-sectional data. For panel data, see [CAUSAL] **xthdidregress**. Below, we illustrate how to use `hdidregress`. For a general overview of DID and more information about the methods used below, see [CAUSAL] **DID intro**. For general discussions about the methods, see Roth et al. (2022) and de Chaisemartin and D'Haultfœuille (2023) and the references therein.

▷ Example 1: Heterogeneous DID for repeated cross-sections

We are interested in knowing if a school-district-level program, Healthy Habits, reduces the body mass index (BMI) for students in the school district. We have fictional data on the Healthy Habits program. This program incorporates more exercise time and augments the intakes of fruits and vegetables. Our data are at the school-district level and include information on whether a school participates in the program, `hhabit`, and the BMI of students in the district, `bmi`. We have repeated samples of students ages 11 to 14 from 40 school districts from the year 2032 to the year 2040.

```
. use https://www.stata-press.com/data/r19/hhabits
(Fictional children BMI and school district data)
```

We are going to use the `aipw` estimator, which allows us to model the outcome and the treatment. If we had selected another estimator and specified the outcome incorrectly, the treatment effects would be inconsistent; see [CAUSAL] **teffects aipw**. With the `aipw` estimator, as long as one of the treatment or outcome model is correctly specified, we will get a consistent estimate of the ATET—a property called double robustness.

We model `hhabit` using the number of parks in the district, `parksd`. We conjecture that school districts with more parks consider exercise spaces more important in their urban planning than those with fewer parks. These districts are therefore more amenable to the Healthy Habits program.

For the outcome variable, we believe that mother's education, `medu`, is a good predictor of the health habits of children. We also believe that participation in sports, `sports`, affects `bmi`. Finally, we control for whether the student is a `girl` to account for behavioral differences and differences in body types of boys and girls at this age.

In the first set of parentheses, we define the outcome, `bmi`, and any covariates that affect the outcome directly. In the second set of parentheses, we define the observation-level treatment variable, `hhabit`, and the covariates that affect it. After the comma, we must define the group variable in `group()`; this is a required option. The group variable defines at which level the treatment occurs and also identifies the clustering variable, which in this case is `schools`. We also need to specify a time variable in `time()`. We fit the following model:

```
. hdidregress aipw (bmi medu i.girl i.sports) (hhabit parksd),
> group(schools) time(year)
note: variable _did_cohort, containing cohort indicators formed by treatment
      variable hhabit and group variable schools, was added to the dataset
      using the estimation sample.

Computing ATET for each cohort and time:
Cohort 2034 (8): ........ done
Cohort 2036 (8): ........ done
Cohort 2038 (8): ........ done
```

Treatment and time information

Time variable: year
Time interval: 2032 to 2040
Control: _did_cohort = 0
Treatment: _did_cohort > 0

	_did_cohort
Number of cohorts	4
Number of obs	
Never treated	11355
2034	1231
2036	2097
2038	2042

Heterogeneous treatment-effects regression — Number of obs = 16,725
Estimator: Augmented IPW
Treatment level: schools
Control group: Never treated

(Std. err. adjusted for 40 clusters in schools)

Cohort	ATET	Robust std. err.	z	P>|z|	[95% conf.	interval]
2034						
year						
2033	.6544681	.5946048	1.10	0.271	-.5109359	1.819872
2034	-1.226451	.379168	-3.23	0.001	-1.969607	-.4832957
2035	-2.491842	.4169657	-5.98	0.000	-3.30908	-1.674605
2036	-2.72486	.2363878	-11.53	0.000	-3.188171	-2.261548
2037	-2.786634	.6672867	-4.18	0.000	-4.094492	-1.478776
2038	-3.980456	.2993279	-13.30	0.000	-4.567127	-3.393784
2039	-.604415	.5929199	-1.02	0.308	-1.766517	.5576866
2040	-.6522272	.3640416	-1.79	0.073	-1.365736	.0612812
2036						
year						
2033	.6635794	.3089663	2.15	0.032	.0580167	1.269142
2034	-1.3933	.3871204	-3.60	0.000	-2.152042	-.6345582
2035	.5947865	.4065947	1.46	0.144	-.2021245	1.391697
2036	-1.71427	.4565384	-3.75	0.000	-2.609069	-.8194714
2037	-3.170542	.5221368	-6.07	0.000	-4.193912	-2.147173
2038	-2.967701	.4247053	-6.99	0.000	-3.800108	-2.135294
2039	.0360098	.6868764	0.05	0.958	-1.310243	1.382263
2040	-.957117	.3510986	-2.73	0.006	-1.645258	-.2689763
2038						
year						
2033	-1.434451	.5163232	-2.78	0.005	-2.446426	-.422476
2034	1.010288	.4808165	2.10	0.036	.067905	1.952671
2035	-.3809733	.4336764	-0.88	0.380	-1.230963	.4690169
2036	.5199519	.4849723	1.07	0.284	-.4305763	1.47048
2037	-.0315794	.5863875	-0.05	0.957	-1.180878	1.117719
2038	-3.602114	.3498692	-10.30	0.000	-4.287845	-2.916383
2039	-1.388906	.6765493	-2.05	0.040	-2.714919	-.0628943
2040	-.6222491	.5510466	-1.13	0.259	-1.70228	.4577824

Note: ATET computed using covariates.
Note: Base time for pretreatment ATETs is the previous period.

Notice the note below the command. A variable with the name `_did_cohort` has been generated. Using the group variable and the observation-level treatment, `hdidregress` generated treatment-time cohorts. The new variable creates treatment groups based on the time when a group was first treated. For instance, if two schools adopt the Healthy Habits program in 2034, they are grouped in the 2034 cohort. The variable also contains a category for a control group. In this case, the control group is formed by the schools that never participate in the program. Cohorts are an important input for estimation and for postestimation commands. You do not need to adhere to the default name, `_did_cohort`, and may provide your own name using the `cohortvar()` option.

Next appears a table that gives you a sense of the treatment groups and time. You see the time variable, `year`, and its range, `2032 to 2040`. Then we see what defines a treated or a control group. The table after provides group-level information about the cohort-time groups. The first row tells you the number of cohorts. Following the number of cohorts is a tabulation showing how many observations are in each cohort. For instance, 11,355 observations are never treated in our data. The table gives you a sense of the amount of information available in each cohort and might hint at the variability of cohort-level estimates.

The next table presents the ATET estimates. The first panel shows the ATETs for the 2034 cohort. We first have the 2033 ATET of 0.65, and the confidence interval includes 0. This is as expected; before treatment, the effect should be 0. We should interpret the ATET to mean that among the school districts that adopted the Healthy Habits program in 2034, the expected `bmi` is 0.65 higher than if the districts had never participated in the program. At treatment onset, in 2034, we observe a treatment effect is a decrease of the `bmi` of 1.23. In the last two periods, the effect of the treatment has diminished for the 2034 cohort; the confidence intervals for the effects in 2039 and 2040 again include 0. We interpret the results for the other cohorts similarly.

◁

▷ Example 2: Visualizing estimation results

In the example above, we had four cohorts and nine time periods. There is a lot of information to process, and it can get even more daunting if we had more cohorts and time periods. To better visualize the results, we can use `estat atetplot`:

```
. estat atetplot
```

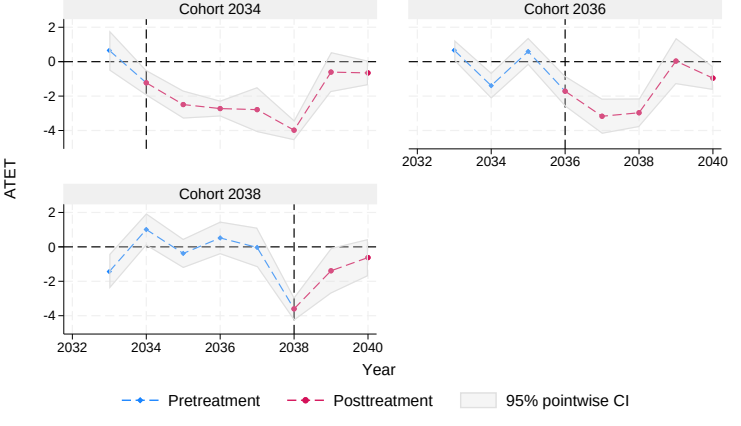

Figure 1. ATETs by cohort over time

The graph shows the pretreatment and the posttreatment ATETs for each cohort and their pointwise confidence intervals. For the 2034 cohort, we see that the program reduces bmi by approximately 2 to 4 points but this tendency seems to start reverting in 2038. A similar pattern emerges over the other two cohorts.

◁

▷ Example 3: Less heterogeneity; aggregating and summarizing treatment effects

So far, we have allowed treatment effects to change over cohort and over time. But we might want to obtain only one treatment effect for each cohort, abstracting away from time variation within cohorts. You would get this using the postestimation command estat aggregation.

```
. estat aggregation, cohort
ATET over cohort                                        Number of obs = 16,725
                              (Std. err. adjusted for 40 clusters in schools)
```

Cohort	ATET	Robust std. err.	z	P>|z|	[95% conf.	interval]
2034	-2.065755	.1999412	-10.33	0.000	-2.457633	-1.673877
2036	-1.7781	.4013978	-4.43	0.000	-2.564825	-.9913744
2038	-1.869405	.4650349	-4.02	0.000	-2.780857	-.9579538

```
Note: Aggregation weights vary across times and cohorts.
```

Note that aggregation occurs only for the posttreatment periods and not for the pretreatment periods. The 2034 estimate is a weighted average of all the treatment-effect estimates after 2034 for the 2034 cohort; see [CAUSAL] **hdidregress postestimation** for more details.

Aggregated estimates are easier to digest; now we have 3 treatment effects to analyze instead of 24. For the 2034 cohort, we have a treatment effect of −2.1. For the 2036 cohort, the effect is −1.8, and for the 2038 cohort, it is −1.9. We cannot see how the treatment evolves over time for each cohort, but we have a sense of the average effect over time for each of them.

We could instead want to see the treatment effect at each point in time, abstracting from cohort-level variation.

```
. estat aggregation, time
ATET over time                                          Number of obs = 16,725
                              (Std. err. adjusted for 40 clusters in schools)
```

Time	ATET	Robust std. err.	z	P>|z|	[95% conf.	interval]
2034	-1.226451	.379168	-3.23	0.001	-1.969607	-.4832957
2035	-2.491842	.4169657	-5.98	0.000	-3.30908	-1.674605
2036	-2.111619	.3654785	-5.78	0.000	-2.827943	-1.395294
2037	-3.028686	.4278557	-7.08	0.000	-3.867268	-2.190104
2038	-3.449829	.2670184	-12.92	0.000	-3.973176	-2.926483
2039	-.6624494	.44865	-1.48	0.140	-1.541787	.2168884
2040	-.7575068	.2816374	-2.69	0.007	-1.309506	-.2055078

```
Note: Aggregation weights vary across times and cohorts.
```

We see the treatment effects for each one of the posttreatment periods. As before, we have the option to look at the effects graphically. We just need to use the `graph` option.

```
. estat aggregation, time graph
```

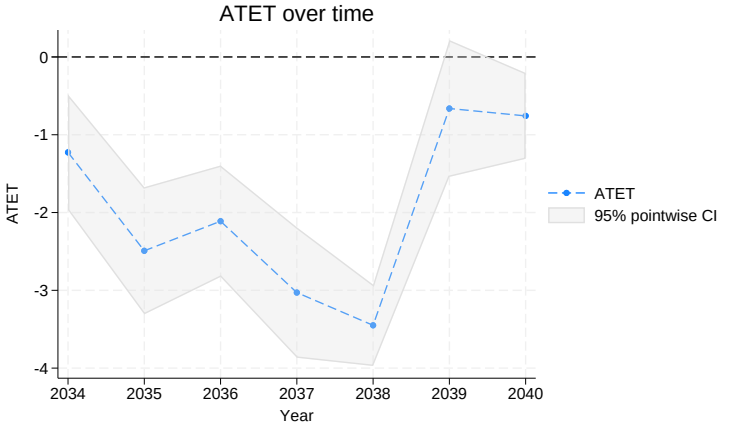

Figure 2. ATETs over time

Example 4: Dynamic treatment effects

We could also ask what the evolution of the treatment effect is after treatment. For instance, we might want to know what happens one period after the onset of treatment, two periods after treatment, and so forth. It might be the case that treatment effects vanish over time or even change patterns. We might also want to see whether, before treatment, we observe a treatment effect or a pattern that might suggest that there is anticipation of treatment. `estat aggregation` allows us to answer these questions by using the `dynamic` option.

```
. estat aggregation, dynamic graph
Duration of exposure ATET                          Number of obs = 16,725
                                (Std. err. adjusted for 40 clusters in schools)
```

Exposure	ATET	Robust std. err.	z	P>|z|	[95% conf.	interval]
-5	-1.434451	.5163232	-2.78	0.005	-2.446426	-.422476
-4	1.010288	.4808165	2.10	0.036	.067905	1.952671
-3	.1338267	.3091619	0.43	0.665	-.4721195	.739773
-2	-.4256324	.4292553	-0.99	0.321	-1.266957	.4156925
-1	.3727141	.3197563	1.17	0.244	-.2539967	.999425
0	-2.285098	.3827362	-5.97	0.000	-3.035248	-1.534949
1	-2.344265	.3829047	-6.12	0.000	-3.094744	-1.593785
2	-2.045521	.3911543	-5.23	0.000	-2.81217	-1.278873
3	-1.045601	.6840119	-1.53	0.126	-2.38624	.2950372
4	-2.145004	.5952525	-3.60	0.000	-3.311678	-.978331
5	-.604415	.5929199	-1.02	0.308	-1.766517	.5576866
6	-.6522272	.3640416	-1.79	0.073	-1.365736	.0612812

```
Note: Base time for pretreatment ATETs is the previous period.
Note: Exposure is the number of periods since the first treatment time.
Note: Aggregation weights vary across times and cohorts.
```

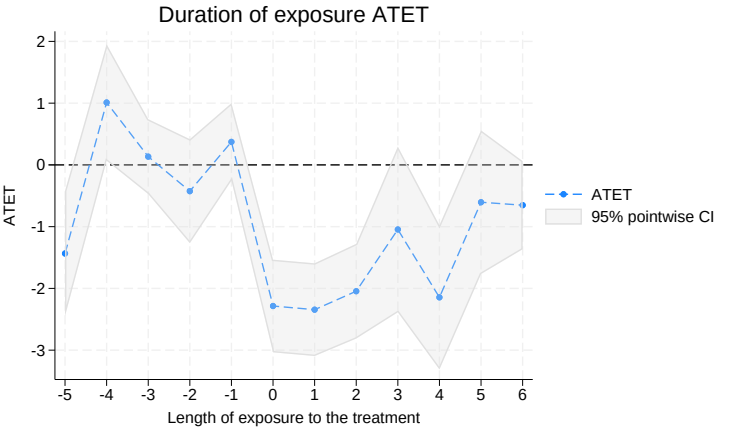

Figure 3. ATET dynamics

In the three periods prior to treatment, there is no effect. This suggests no anticipation to treatment. At the onset, the program reduces `bmi`, but the effect decreases for school districts that remain for more than four years in the program.

◁

▷ Example 5: TWFE estimation

The literature on heterogeneous DID started by pointing out the problems that arise when one assumes erroneously that the treatment effects are homogeneous. It suggested that TWFE estimation was inadequate. Wooldridge (2021) suggests that fixed-effects estimation can be used if we extend it to include interactions between treatment-time cohorts and time.

Another important insight of Wooldridge (2021) is that you can use pooled ordinary least squares and add panel-level averages of covariates and obtain the same point estimates as one would get with fixed-effects estimation in the context of DID estimation. This is an extension of the intuition by Mundlak (1978). `xthdidregress` and `hdidregress` fit pooled ordinary least-squares models using these ideas. Below, we present the results we obtain using the `twfe` estimator.

```
. hdidregress twfe (bmi medu i.girl i.sports) (hhabit), group(schools) time(year)
note: variable _did_cohort, containing cohort indicators formed by treatment
      variable hhabit and group variable schools, was added to the dataset
      using the estimation sample.

Treatment and time information

Time variable: year
Time interval: 2032 to 2040
Control:       _did_cohort = 0
Treatment:     _did_cohort > 0
```

	_did_cohort
Number of cohorts	4
Number of obs	
Never treated	11355
2034	1231
2036	2097
2038	2042

```
Heterogeneous treatment-effects regression              Number of obs = 16,725
Data type:       Repeated cross-sectional
Estimator:       Two-way fixed effects
Treatment level: schools
Control group:   Never treated
Heterogeneity:   Cohort and time
```

(Std. err. adjusted for 40 clusters in schools)

Cohort	ATET	Robust std. err.	t	P>|t|	[95% conf.	interval]
2034						
year						
2034	-.8057824	.2723491	-2.96	0.005	-1.35666	-.2549045
2035	-1.951481	.2098279	-9.30	0.000	-2.375898	-1.527064
2036	-2.091438	.2081903	-10.05	0.000	-2.512542	-1.670333
2037	-2.329408	.4674253	-4.98	0.000	-3.274865	-1.383952
2038	-3.623645	.4658056	-7.78	0.000	-4.565826	-2.681464
2039	-.1729334	.7543583	-0.23	0.820	-1.698767	1.3529
2040	-.2267266	.3344035	-0.68	0.502	-.9031216	.4496684
2036						
year						
2036	-1.671963	.3424563	-4.88	0.000	-2.364646	-.9792798
2037	-3.27542	.3496365	-9.37	0.000	-3.982627	-2.568213
2038	-2.995124	.2853544	-10.50	0.000	-3.572308	-2.41794
2039	-.0792949	.5152787	-0.15	0.878	-1.121544	.9629547
2040	-.9852905	.1856743	-5.31	0.000	-1.360852	-.6097289
2038						
year						
2038	-3.389082	.154181	-21.98	0.000	-3.700942	-3.077221
2039	-.7309226	.5173441	-1.41	0.166	-1.77735	.3155046
2040	-.6942153	.3558485	-1.95	0.058	-1.413987	.0255563

Note: ATET computed using covariates.

The output is almost the same as the one for the ra estimator in example 1. There are a couple of noteworthy differences. First, the estimator fits an extended TWFE regression. Second, the ATET parameters are shown for each cohort only at the time of treatment exposure and for the periods thereafter but not for the pretreatment periods. As discussed in Wooldridge (2021), these are the parameters identified using the parallel-trends assumption he derives.

As we did before, we could use estat aggregation to explore different ways of looking at our treatment effects and estat atetplot to visualize the estimated ATETs.

◁

▷ Example 6: Reducing model complexity

When we fit the aipw model, we had to estimate ATET parameters for each cohort over time. The complexity of the model grows with the number of cohorts and the number of time periods. As is described in *Methods and formulas*, the aipw estimator uses a different subset of the data to obtain each parameter. To get a reliable estimator of each parameter, you need sufficient data for each subsample. Sometimes, there are few observations for a given cohort in a given set of time periods.

We can ameliorate this problem by reducing the amount of heterogeneity we assume. For the twfe estimator, the complexity of the model comes from the interactions between the observation-level treatment with cohort and time and the interactions between the observation-level treatment, cohort, time, and covariates. This allows us to decide which interactions to include in our model. We could, for instance, allow for heterogeneity at the cohort level instead of at the cohort and time level. We use the hettype() option with the argument cohort to do this:

```
. hdidregress twfe (bmi medu i.girl i.sports) (hhabit), group(schools)
> time(year) hettype(cohort)
note: variable _did_cohort, containing cohort indicators formed by treatment
      variable hhabit and group variable schools, was added to the dataset
      using the estimation sample.

Treatment and time information

Time variable: year
Time interval: 2032 to 2040
Control:       _did_cohort = 0
Treatment:     _did_cohort > 0
```

	_did_cohort
Number of cohorts	4
Number of obs	
Never treated	11355
2034	1231
2036	2097
2038	2042

```
Heterogeneous treatment-effects regression              Number of obs = 16,725
Data type:       Repeated cross-sectional
Estimator:       Two-way fixed effects
Treatment level: schools
Control group:   Never treated
Heterogeneity:   Cohort
                                (Std. err. adjusted for 40 clusters in schools)
```

Cohort	ATET	Robust std. err.	t	P>|t|	[95% conf.	interval]
2034	-1.619553	.2223114	-7.29	0.000	-2.069221	-1.169886
2036	-1.832602	.1954433	-9.38	0.000	-2.227924	-1.437281
2038	-1.739144	.2152765	-8.08	0.000	-2.174582	-1.303706

```
Note: ATET computed using covariates.
```

You fit a regression model with fewer terms and obtain treatment effects only at the cohort level. You could also have the treatment effect change over time but not over cohort by typing `hettype(time)`.

For the estimators proposed by Callaway and Sant'Anna (2021), heterogeneity is built in, so we need to estimate all the ATET parameters.

◁

▷ Example 7: Defining your own cohort

By default, `hdidregress` creates a cohort variable based on the estimation sample. Yet this might be inadequate if a researcher has more information than is provided in the dataset. Suppose that our dataset looked something like this for school district 1:

```
. list schools year hhabit in 100/105, noobs sepby(schools)
```

schools	year	hhabit
1	2033	No
1	2033	No
1	2035	Yes
1	2035	Yes
1	2035	Yes
1	2035	Yes

There is no information for the year 2034. If the school district participated in the healthy habits program in 2034, it should belong to the 2034 cohort. However, `hdidregress` has no information about the year 2034 in the estimation sample and will classify school district 1 as belonging to the 2035 cohort. `hdidregress`'s inability to determine the proper cohort is not exclusive to situations with gaps in your repeated cross-section. In fact, Stata excludes observations in your sample if any of the variables used during estimation are missing. If all observations for the time period in which a group is first treated are omitted because of missing values, `hdidregress` cannot assign the group to the appropriate cohort.

If you have information about the cohort values, instead of letting the command create a cohort variable, you can provide the cohort variable with the `usercohort()` option. Suppose you had a cohort variable, `mycohort`; then you could type

```
. hdidregress twfe (bmi) (hhabit), group(schools) time(year) usercohort(mycohort)
```

Another possibility is to generate the cohort variable mycohort yourself using the gencohort command; this is helpful when you have missing information on covariates or the outcome but have enough information about the treatment. Suppose you had missing information about the outcome variable bmi but had information about the treatment variable. Below, we drop information about the 2034 cohort to illustrate the point.

```
. replace bmi = . if year==2034 & schools==1
(44 real changes made, 44 to missing)
```

These observations for year 2034 would not be used during estimation, but we have enough information in them to create our own cohort variable.

```
. gencohort mycohort, treat(hhabit) time(year) group(schools)
. list schools year hhabit bmi mycohort in 100/105, noobs sepby(schools)
```

schools	year	hhabit	bmi	mycohort
1	2033	No	20.14775	2034
1	2033	No	21.06941	2034
1	2034	Yes	.	2034
1	2034	Yes	.	2034
1	2034	Yes	.	2034
1	2034	Yes	.	2034

The mycohort variable can now be specified in the usercohort() option of hdidregress() to properly treat school district 1 as belonging to cohort 2034.

◁

Stored results

hdidregress stores the following in e():

Scalars

e(N)	number of observations
e(N_clust)	number of clusters
e(tmin)	first time period
e(tmax)	last time period
e(rank)	rank of e(V)

Macros

e(cmd)	hdidregress
e(cmdline)	command as typed
e(clustvar)	name of cluster variable
e(control_group)	control group
e(het_type)	heterogeneity type for twfe estimator
e(cohortvar)	name of cohort variable
e(usercohort)	name of user-specified cohort variable
e(ovar)	name of outcome variable
e(wtype)	weight type
e(wexp)	weight expression
e(marginsnotok)	predictions disallowed by margins
e(timevar)	time variable
e(treatname)	name of treatment variable
e(basetime)	type of pretreatment base time
e(estat_cmd)	program used to implement estat
e(vce)	*vcetype* specified in vce()
e(vcetype)	title used to label Std. err.

e(method)	estimator method
e(properties)	b V
Matrices	
e(b)	coefficient vector
e(V)	variance–covariance matrix of the estimators
e(cohort_count)	matrix with cohort count information
Functions	
e(sample)	marks estimation sample

In addition to the above, the following is stored in r():

Matrices	
r(table)	matrix containing the coefficients with their standard errors, test statistics, p-values, and confidence intervals

Note that results stored in r() are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

Methods and formulas

Methods and formulas are presented under the following headings:

Introduction

hdidregress for repeated cross-sectional data implements the RA, IPW, and AIPW estimators, outlined in Callaway and Sant'Anna (2021), and the TWFE estimator, outlined in Wooldridge (2021).

To reveal how the heterogeneous treatment effects evolve across cohorts and time, we are interested in estimating the ATET for each combination of cohort and time. Cohorts are defined by the time a group is treated, where time is denoted by t, where $t = 1, \ldots, T$. We denote a cohort by g and the individuals in our sample by i, where $i = 1, \ldots, N$. Let G_{ig} be an indicator that equals one if unit i is first treated at time g. Then the units in cohort g can be denoted by $G_{ig} = 1$. When a unit i is never treated, we denote $G_{i0} = 1$. Thus, cohort 0 indicates all the units that are never treated. We assume that once a unit is treated, it will remain treated. We also define d_{it} as an indicator for treatment of unit i at time t.

Let $\theta(g,t)$ be the ATET for cohort g at time t, which is defined as

$$\theta(g,t) = \mathbf{E}\left\{y_t(g) - y_t(0) | G_g = 1\right\} \qquad \text{(ATET)}$$

where $y_t(g)$ is the potential outcome at time t for those first treated at time g, $y_t(0)$ is the potential outcome for those that are never treated, and G_g equals 1 if a unit belongs to cohort g. All the four estimators provided in hdidregress estimate $\theta(g,t)$ in equation (ATET). We cannot directly estimate $\theta(g,t)$ using equation (ATET) because the potential outcomes $y_t(g)$ and $y_t(0)$ are not observable.

Next, we will describe the RA, IPW, and AIPW estimators.

The RA, IPW, and AIPW estimators

To estimate the ATET for cohort g at time t, the RA, IPW, and AIPW estimators transform the estimation into a classical two groups and two periods difference-in-differences setup. Thus, we need to restrict the data to an estimation sample with only two groups and only two periods based on the values of g and t. For the two groups, one group comprises all observations in cohort g; the other group comprises untreated observations not in cohort g, also known as a control group. For the two periods, one period is the data in time t; the other period is a period when cohort g is not treated, also known as base time.

There are two ways to define the control group. One way is to use the units that are never treated as the control group. Let C^{NEV} be an indicator that equals one if a unit belongs to the never-treated group. In particular, $C^{\text{NEV}} = G_0$. Another way is to use the units not in cohort g and not yet treated at time t as the control group. Let $C^{\text{NY}}_{g,t}$ be an indicator that equals one if a unit belongs to the not-yet-treated group by time t. In particular, $C^{\text{NY}}_{g,t} = (1 - G_g)(1 - d_t)$. To simplify, we indicate control, in both cases, as $C^*_{g,t}$.

The definitions of the RA, IPW, and AIPW estimators depend on the definition of $C^*_{g,t}$, which can either be C^{NEV} or $C^{\text{NY}}_{g,t}$. However, regardless of the control group's choice, the estimators' definitions can always be written using the general notation $C^*_{g,t}$.

There are also two ways to define the base time. One way is to adaptively choose the base time for the pretreatment periods. When the adaptive method is used to compute the ATET for cohort g at time t, for the pretreatment periods, the base time is $t - 1$; for the posttreatment periods, the base time is $g - 1$. Another way is to use a common base time $g - 1$ for both pretreatment and posttreatment periods. The common base time is useful for identifying a violation of the parallel trends assumption in event studies as discussed in Roth (2024). To simplify the notation, we indicate the base time in both cases as t_0.

For each unit i in the pooled sample, we observe $\{\tau_i, y_{i,\tau_i}, \mathbf{x}_{i,\tau_i}, d_{i,\tau_i}, \mathbf{z}_{i,\tau_i}\}$, where y_i is the outcome, $\mathbf{x}_i$ are pretreatment covariates for the outcome model, d_i is a treatment indicator, $\mathbf{z}_i$ are covariates for the treatment assignment model, and $\tau_i \in \{1, \ldots, T\}$ is a categorical variable indicating the time when unit i is observed. Let T_t equal one if the unit is observed at time t and zero otherwise.

The estimands also require the following notation,

$$
\begin{aligned}
m^{\text{treat}}_{g,s}(\mathbf{x}) &= \mathbf{E}(y|\mathbf{x}, G_g = 1, \tau = s) \\
m^{\text{comp}}_{g,s,t}(\mathbf{x}) &= \mathbf{E}(y|\mathbf{x}, C^*_{g,t} = 1, \tau = s) \\
w^{\text{treat}}_{g,s} &= \frac{T_s G_g}{\mathbf{E}(T_s G_g)} \\
w^{\text{comp}}_{g,s,t}(\mathbf{z}) &= \frac{\frac{T_s p_{g,t}(\mathbf{z}) C^*_{g,t}}{1 - p_{g,t}(\mathbf{z})}}{\mathbf{E}\left\{\frac{T_s p_{g,t}(\mathbf{z}) C^*_{g,t}}{1 - p_{g,t}(\mathbf{z})}\right\}}
\end{aligned}
$$

where $p_{g,t}(\mathbf{z})$ is defined by

$$
p_{g,t}(\mathbf{z}) = \Pr(G_g = 1|\mathbf{z}, G_g + C^*_{g,t} = 1) \tag{Pz}
$$

and the superscript refers to the group we are conditioning on, either the treated group (treat) or the control or comparison group (comp).

The RA estimand is

$$\theta_{\text{RA}}(g,t) = \mathbf{E}\left(\frac{G_g}{\mathbf{E}(G_g)}\left[\{m^{\text{treat}}_{g,t}(\mathbf{x}) - m^{\text{treat}}_{g,g-1}(\mathbf{x})\} - \{m^{\text{comp}}_{g,t,t}(\mathbf{x}) - m^{\text{comp}}_{g,g-1,t}(\mathbf{x})\}\right]\right) \tag{RA}$$

The IPW estimand is

$$\theta_{\text{IPW}}(g,t) = \mathbf{E}\left\{\left(w^{\text{treat}}_{g,t} - w^{\text{treat}}_{g,g-1}\right)y\right\} - \mathbf{E}\left[\{w^{\text{comp}}_{g,t,t}(\mathbf{z}) - w^{\text{comp}}_{g,g-1,t}(\mathbf{z})\}y\right] \tag{IPW}$$

The AIPW estimand is

$$\begin{aligned}\theta_{\text{AIPW}}(g,t) = {}& \mathbf{E}\left(\frac{G_g}{\mathbf{E}(G_g)}\left[\{m^{\text{treat}}_{g,t}(\mathbf{x}) - m^{\text{treat}}_{g,g-1}(\mathbf{x})\} - \{m^{\text{comp}}_{g,t,t}(\mathbf{x}) - m^{\text{comp}}_{g,g-1,t}(\mathbf{x})\}\right]\right)\\ &+ \mathbf{E}\left[w^{\text{treat}}_{g,t}\{y - m^{\text{treat}}_{g,t}(\mathbf{x})\} - w^{\text{treat}}_{g,g-1}\{y - m^{\text{treat}}_{g,g-1}(\mathbf{x})\}\right]\\ &- \mathbf{E}\left[w^{\text{comp}}_{g,t,t}(\mathbf{z})\{y - m^{\text{comp}}_{g,t,t}(\mathbf{x})\} - w^{\text{comp}}_{g,g-1,t}(\mathbf{z})\{y - m^{\text{comp}}_{g,g-1,t}(\mathbf{x})\}\right]\end{aligned} \tag{AIPW}$$

Under some regularity conditions, Callaway and Sant'Anna (2021) showed that the estimand for RA, IPW, and AIPW is the same as $\theta(g,t)$ in equation (ATET). In other words,

$$\theta(g,t) = \theta_{\text{RA}}(g,t) = \theta_{\text{IPW}}(g,t) = \theta_{\text{AIPW}}(g,t)$$

Furthermore, the estimands in equations (AIPW) are estimable because they are all based on observed variables. The identification of the estimators sheds light on how to estimate $\theta(g,t)$. The estimator can be generally divided into three steps:

1. Restrict the sample to time t and t_0, and keep only the units in cohort g or in control group $C^*_{g,t}$. When option `basetime(adaptive)` is specified, $t_0 = g-1$ if $t \geq g$ or $t_0 = t-1$ if $t < g$. When option `basetime(common)` is specified, $t_0 = g-1$.

2. Use a parametric model to estimate the nuisance functions.

   a. For outcomes: linear regression to estimate $m^{\text{treat}}_{g,t}(\mathbf{x})$, $m^{\text{treat}}_{g,t_0}(\mathbf{x})$, $m^{\text{comp}}_{g,s,t}(\mathbf{x})$, and $m^{\text{comp}}_{g,s,t_0}(\mathbf{x})$.

   b. For propensity: logit regression to estimate $p_{g,t}(\mathbf{z})$.

   c. For probability weights: $w^{\text{treat}}_{g,t}$, w^{treat}_{g,t_0}, $w^{\text{comp}}_{g,s,t}(\mathbf{z})$, and $w^{\text{comp}}_{g,s,t_0}(\mathbf{z})$ to estimate using propensity scores T_t and G_g.

3. Plug in the nuisance function estimates into the estimating equation in equations (IPW), (AIPW), or (RA). Notice that the expectation operator $\mathbf{E}(\cdot)$ is replaced by the sample average.

The variance–covariance matrix for the estimates is computed using the influence-function approach proposed in Callaway and Sant'Anna (2021). The influence function approach is numerically equivalent to the generalized method of moments approach. However, it is much faster because it avoids computing the covariance matrix for the parameters in the nuisance functions. For more discussions on influence functions, see Hampel et al. (1986), Newey and McFadden (1994), and Jann (2020).

The TWFE estimator

A TWFE estimator for repeated cross-sections fits

$$y_i = \alpha_h + \gamma_t + \mathbf{x}_i\boldsymbol{\beta} + d_i\tau + \epsilon_i$$

Above, h denotes the group level at which treatment occurs. Wooldridge (2021) extends this model to incorporate interactions between the observation-level treatment, d_i, cohort, G_{ig}, posttreatment periods, and covariates. We define indicators for the posttreatment period as f_s with s going from q to T, where q is the first time period we observe treatment. For instance, f_q equals 1 if we are in time-period q and 0 otherwise. To simplify this notation, we show the model without covariates. The extended fixed-effects model is given by

$$y_i = \eta + \sum_{g=q}^{T} G_{ig}\theta_g + \sum_{s=2}^{T} f_s\gamma_s + \sum_{g=q}^{T}\sum_{s=g}^{T} d_i G_{ig} f_s \tau_{gs} + \epsilon_i \qquad \text{(TWFE)}$$

We can fit equation (TWFE) using pooled ordinary least squares or a within estimator. We are going to use the estimator proposed by Mundlak (1978). This gives the same point estimates as using the within estimator with h as the panel level for the parameters in equation (TWFE) but has different degrees of freedom because of the additional terms added by the Mundlak approach. Unlike within estimation, the Mundlak approach works for both repeated cross-sectional data as well as for panel data. Also, it has good properties to obtain partial effect under various data-generating processes, as pointed out in Wooldridge (2019).

Above, the τ_{gs} are the cohort-time treatment effects. When we have covariates, we interact them with all the relevant variables in the model. To get the treatment effects in this case, we need to control for the variation in the covariates. We can obtain both effects using `margins` by typing

```
. margins, dydx(d) at(year=q ... year=T) over(cohort) vce(unconditional)
```

where `d` is the treatment indicator, `year` indicates treatment times at which treatment will be evaluated using `at()`, and `cohort` is the treatment-time cohorts. We use `vce(unconditional)` to account for the variation in the covariates.

In practice, `hdidregress` computes the treatment effects analytically rather than by use of `margins`. Specifically, a modified Mundlak regression model is fit. The modified regression interacts treatment indicators with covariates demeaned by cohort-specific means instead of the covariates themselves. Treatment-effect parameters can be estimated as coefficients of this regression rather than as linear combinations of regression coefficients, even when covariates are present.

The modified Mundlak regression is treated as being fit following a set of first-stage regressions of each covariate on cohort indicators. GMM-style standard errors account for variation in these first-stage regressions and are equivalent to the standard errors produced by `margins` with the `vce(unconditional)` option.

With the `hettype()` option, we reduce the complexity of (TWFE). In particular, if we ask for `hettype(time)`, we have

$$y_i = \eta + \sum_{g=q}^{T} G_{ig}\theta_g + \sum_{s=2}^{T} f_s\gamma_s + \sum_{s=q}^{T} d_i f_s \tau_s + \epsilon_i$$

Now treatment varies over time but not over cohort, that is, τ_s. If we use the `hettype(cohort)` option, we have

$$y_i = \eta + \sum_{g=q}^{T} G_{ig}\theta_g + \sum_{s=2}^{T} f_s\gamma_s + \sum_{g=q}^{T} d_i G_{ig}\tau_g + \epsilon_{it}$$

Now treatment varies over cohort but not over time, that is, τ_g.

When the `controlgroup(notyet)` option is specified, the G_{ig} indicator excludes the last treated cohort. As discussed in Wooldridge (2021), when every group is eventually treated, we cannot identify the treatment effect for this cohort. It is therefore sensible to use the last treated cohort as a control group. When some of the units in our sample are never treated, we can always identify all cohorts, and the `twfe` estimator will always revert to using `controlgroup(never)`.

Acknowledgments

We thank Fernando Rios-Avila of the Levy Economics Institute of Bard College for helpful conversations and advice and for his contributions in developing heterogeneous treatment-effects software and accompanying material. His community-contributed commands `csdid` and `drdid` implement some of the methods in this entry and a few others we have yet to include.

We thank Pedro H. C. Sant'Anna of Emory University for helpful conversations and advice and for his continuing contributions to the heterogeneous treatment-effects literature.

We thank Jeff Wooldridge of Michigan State University for helpful conversations about extended two-way fixed-effects estimation.

We also thank Austin Nichols of Amazon for his advice, for his insights, and for organizing the WHYDC causal inference conference.

Stata has an active research community adding features to the area of heterogeneous DID. We would like to acknowledge their previous and ongoing contributions to this area: `csdid` by Fernando Rios-Avila; `did2s` by Kyle Butts; `did_multiplegt` by Clement de Chaisemartin and Xavier D'Haultfœuille; `did_imputation` by Kirill Borusyak, Xavier Jaravel, and Jann Spiess; `eventstudyinteract` by Liyang Sun and Sarah Abraham; `jwdid` by Fernando Rios-Avila; and many more. Type `search heterogeneous difference in differences` to see Stata's official and community-contributed features for heterogeneous DID.

References

Callaway, B., and P. H. C. Sant'Anna. 2021. Difference-in-differences with multiple time periods. *Journal of Econometrics* 225: 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001.

de Chaisemartin, C., and X. D'Haultfœuille. 2023. Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *Econometrics Journal* 26: C1–C30. https://doi.org/10.1093/ectj/utac017.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions.* New York: Wiley. https://doi.org/10.1002/9781118186435.

Jann, B. 2020. Influence functions continued. A framework for estimating standard errors in reweighting, matching, and regression adjustment. Working Papers 35, University of Bern Social Sciences. https://ideas.repec.org/p/bss/wpaper/35.html.

Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica* 46: 69–85. https://doi.org/10.2307/1913646.

Newey, W. K., and D. L. McFadden. 1994. "Large sample estimation and hypothesis testing". In *Handbook of Econometrics*, edited by R. F. Engle and D. L. McFadden, vol. 4: 2111–2245. Amsterdam: Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4.

Roth, J. 2024. Interpreting event-studies from recent difference-in-differences methods. ArXiv Working Paper No. arXiv:2401.12309, https://doi.org/10.48550/arXiv.2401.12309.

Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe. 2022. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. arXiv:2201.01194 [econ.EM], https://doi.org/10.48550/arXiv.2201.01194.

Wooldridge, J. M. 2019. Correlated random effects models with unbalanced panels. *Journal of Econometrics* 211: 137–150. https://doi.org/10.1016/j.jeconom.2018.12.010.

———. 2021. Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. Working paper, Department of Economics, Michigan State University, East Lansing, MI. https://doi.org/10.2139/ssrn.3906345.

Also see

[CAUSAL] **hdidregress postestimation** — Postestimation tools for hdidregress and xthdidregress

[CAUSAL] **xthdidregress** — Heterogeneous difference in differences for panel data

[CAUSAL] **DID intro** — Introduction to difference-in-differences estimation

[CAUSAL] **didregress** — Difference-in-differences estimation+

[CAUSAL] **gencohort** — Create a cohort variable

[CAUSAL] **teffects intro** — Introduction to treatment effects for observational data

[CAUSAL] **teffects intro advanced** — Advanced introduction to treatment effects for observational data

[U] **20 Estimation and postestimation commands**